DOCUMENT RESUME

ED 334 237 TM 016 762

AUTHOR Sykes, Robert C.; And Others
TITLE Assessing the Effects of Computer Administration on Scores and Parameter Estimates Using IRT Models.
PUB DATE Apr 91
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adults; Certification; Comparative Testing; *Computer Assisted Testing; *Item Response Theory; *Licensing Examinations (Professions); Psychometrics; *Scores; Test Construction; Test Format; Test Items; *Test Validity
IDENTIFIERS *Paper and Pencil Tests

ABSTRACT

To investigate the psychometric feasibility of replacing a paper-and-pencil licensing examination with a computer-administered test, a validity study was conducted. The computer-administered test (Cadm) was a common set of items for all test takers, distinct from computerized adaptive testing, in which test takers receive items appropriate to their estimated abilities. The Cadm version, scheduled for implementation in 1990, would consist of 230 items. The validation study was designed to use the capabilities of item response theory (IRT) to produce shorter test forms of a specified reliability relative to the original test or the proposed version. Four combinations of paper-and-pencil and Cadm tests were administered to 418 licensure candidates, each of whom were administered 150 items each in a paper-and-pencil mode and in a Cadm mode. No effects of the mode of administration were found, and there were no signs of significant differences in overall performance across the administration modes. Assessment of multiple item forms and candidate samples reinforced the conclusion that there was no significant effect of computer administration on candidate performance. Eleven tables and eight figures present study results. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334237

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT C. SYKES

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ASSESSING THE EFFECTS OF COMPUTER ADMINISTRATION ON SCORES AND PARAMETER ESTIMATES USING IRT MODELS

Robert C. Sykes
Helen Heydeman
Mike Whitaker
Doug Lownsbery
Robert Kelly
Billie Haynes

CTB/Macmillan/McGraw-Hill
April 1991

This paper was presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1991.

TM 016 762

BEST COPY AVAILABLE

With the arrival of increasingly powerful, inexpensive microcomputers, the automated administration of educational or occupational tests offers a potentially attractive alternative to conventional paper-and-pencil administration. Automated tests can increase test security, reduce costs of test production and administration, and provide for an immediate report of test results.

Two approaches to automated testing have emerged in educational measurement literature. One of these approaches is computerized adaptive testing (CAT). This approach differs from conventional testing in that each test-taker receives a set of test items that have psychometric properties appropriate to his or her estimated ability level. In conventional testing, a common test is administered to all.

The second approach is an automated administration of a common set of items to all test-takers, as is conducted in traditional paper-and-pencil administrations. This approach hereafter will be referred to as Cadm, short for computer-administered, even though the CAT testing procedure is, by definition, also computer driven.

The potential advantages afforded by automated testing do not mitigate the importance, however, of ensuring that scores from an automated testing procedure can be interpreted in a manner comparable to scores obtained from previous operational testing procedures. The equivalence of scores between modes of administration must be established. This equivalence not only allows a score from a computerized examination to be meaningfully compared to a score from a conventional examination, but also permits a cutscore established for the conventional examination to be applied to the corresponding computer-administered examination.

Establishing the equivalency of scores across different modes of administration requires more than merely verifying that the rank order of scores and score distributions are similar across modes (Guidelines for Computer-Based Tests and Interpretations, APA, 1986). In order to construct test forms that are similar across modes, the generalizability of item statistics or parameters obtained for items administered under previous testing procedures to computer administrations must be assured. The potential for differences in item presentation to affect item parameters from a utilized item response theory model must be investigated and the magnitude of any such effects documented.

For example, if lengthy item text necessitates scrolling that could subsequently cause a computer-administered examination to be more speeded than a conventional examination, then this difference in item presentation would result in increased difficulty of at least some of the computer-administered examination items.

Bunderson et al. (1989) and Mazzeo and Harvey (1988) summarized work on a number of research issues concerning CB versus conventional tests. Bunderson et al. reported that the reliabilities of CB and conventional tests are often very similar. Mazzeo and Harvey, citing studies demonstrating equivalence of scores across administration modes as well as those that do not support score equivalency recommended that equivalency between the two modes not be assumed. Specifically these authors, congruent with a similar recommendation in the APA Guidelines, urged that separate equating and/or norming studies be conducted.

In many of the studies assessing equivalence in the context of achievement and ability testing, the mean score obtained from the conventional test has exceeded, though not always significantly, the mean score produced on the CB test (Wise and Plake, 1989). Such differences, even though small, would be of concern for users of criterion-referenced tests that utilized cutoff scores, originally set for p&p tests, as the basis for licensing decisions. The appropriateness of any p&p cutscore applied to a CB test would be questionable in these circumstances unless these differences could be eliminated through equating.

Mazzeo and Harvey argued, however, that the possibility of asymmetric practice effects which could be larger when one type of exam is administered after the other rather than in the reverse order should preclude the conducting of equating studies based on single-group counterbalanced designs. Furthermore, these authors noted that results of a number of equating studies were difficult to interpret because of the confounding of alternate forms of a test with computer versus conventional administration, or the confounding of order of administration with intervening learning or other factors.

Experimental designs that permit independent tests of mode of administration effects unfortunately make substantial demands, in terms of testing time, on subjects. If the conventional test is very large, the testing time required to take several forms becomes prohibitive. Under these conditions, methods are required to produce reliable part-forms that may be used to make valid inferences about performance on the full p&p and CB forms.

The use of items that have been scaled using IRT permits the construction of shortened parallel conventional and CB forms and ensures that these forms will have a specified reliability relative to the full forms. Experimental designs or validation studies may then be constructed to include several factors, including order of administration. The number of factors that may be assessed as well as the power of significance tests of these effects may be further increased by within-subject designs giving each subject more than one shortened form.

Method

Research Design

In order to investigate the psychometric feasibility of replacing a paper-and-pencil licensure examination with a Cadm testing procedure, a validation study was authorized by a licensing board that is a client of CTB.

The studied licensure examination has traditionally been a paper-and-pencil (p&p) test consisting of 250 real, or scored, items if no items are deleted because they have negative item point biserial coefficients. The p&p examination is administered twice a year, with each examination constructed from an item bank that has been calibrated using the item response theory (IRT) one parameter (Rasch) model. The Cadm version of the examination is scheduled for implementation in 1990. The test will consist of 230 real, or scored, items.

It is desirable when examining the effects of different testing conditions, such as the mode of administration of an examination, to expose each examinee to more than one experimental condition. Such within-subject experimental designs allow a more precise estimate of the condition effect by controlling for variation in examinees' performance. Thus a more precise estimate of the effect of a computer administration of the exam on candidate scores and item parameters could be obtained by giving a sample of candidates both a Cadm and a conventional p&p examination.

However the task of candidates taking a 250 item p&p examination and a set of computer-administered items as large as the 230 planned for a Cadm exam in one setting would be arduous. Consequently, a validation study design was devised that exploited the capabilities of IRT to produce forms that were shorter than either the traditional p&p or the new Cadm forms yet were of a specified reliability relative to these longer forms. These shortened p&p and Cadm forms, each selected to be proportionally representative of the test plan category quotas, would then permit a within-subject candidate validation design that could be administered within a period of time comparable to that permitted for a p&p administration, thus minimizing the possibility of fatigue affecting candidate performance.

The shortened validation p&p and Cadm forms were designed to consist of component item sets of 30 items each. As opposed to a form, which was proportionally test plan representative and had similar psychometric characteristics as past p&p exams, item set content was determined by an approximate balancing of form content over item set constituents. The allocation of item sets to the different test conditions is portrayed in Figure 1. Though not explicitly denoted in the figure, the item sets constituted a total of eight forms with four of the eight composed of two constituent

forms. In order of increasing size and labeled by their item set constituents, the eight forms were: DE, IJ, ABC, FGH, ABCDE, ABCIJ, FGHIJ, and FGHDE.

As previously indicated, each form consisted of items that were proportionally test plan representative and had average item statistics very similar to past p&p exams. Compared to one p&p examination administered in 1989 (reference exam), no form had an average item p value that differed from the reference average p value by more than .02 or an average item point biserial coefficient that differed from the reference average point biserial by more than .05. The fit of the selected items to the one-parameter model was good, with no form having an average Yen's Q1 chi-square statistic (Yen, 1981) differing by more than .84 from the reference average.

Standard error (s.e.) curves for all forms were similar in shape and location, although not necessarily in height, to that for the reference test. Differences in s.e. levels occur predictably with higher s.e.'s obtained for forms with fewer items. The s.e. curves were approximately coincident within each of the following groups of forms: ABCDE, FGHIJ, ABCIJ, FGHDE; ABC and FGH; and DE and IJ. The parallel nature of the forms within each of the above three groups permitted the examination of a number of potential effects of computer administration without having to administer any item more than once to any candidate.

The eight forms were configured into four different combinations of p&p and Cadm item sets in a manner that would allow independent tests of these effects. The four different combinations, called conditions in this study, were defined by two different orders of administration (Cadm first vs. p&p first) crossed with two different positionings of the DE and IJ forms (DE, or IJ, with the ABC or the FGH form).

The positioning of the DE and IJ modules, noted in boldface in Figure 1, permitted powerful tests of the effects of mode of administration on item parameters because of their presence in both computer and p&p forms. In addition, the counterbalancing of administration order and the administration of the ABC form solely in a p&p mode allowed an unconfounded assessment of differences in item parameters or scores due to administering a given form of items before or after administering 150 other items. If an order effect was present, the ABC form afforded an estimate of the magnitude of the effect. Additionally, scores on the ABC form permitted a partial assessment of the equivalence of the four samples of candidates assigned to the four different experimental conditions.

The FGH form was similar to the ABC form in that the FGH form of 90 items was always administered in the same mode (Cadm as opposed to p&p, however) but differed from the ABC form in one important

respect. The items in the G set of items in Conditions 3 and 4, denoted by an asterisk in Figure 1, were administered in a random order as opposed to the fixed order of Conditions 1 and 2. A random sequencing of items, with the added constraint of not allowing the most difficult items to be the first presented to a candidate, is a feature tentatively planned for operational Cadm administrations. The placement of the random item location G set of items allowed an independent estimate of any item allocation effect even in the presence of any previously determined order of administration effect.

The Sample

A total of 418 licensure candidates were recruited to participate in the validation study. These candidates reflected a mix of first-timers and repeaters (80% versus 20%) that was representative of the percentage of repeaters participating in the last several p&p exams. First-time candidates were recruited from accredited schools and were recent graduates or within the last month of their program. An incentive in the form of an examination fee waiver was offered to each candidate who completed both p&p and Cadm tests.

Test Administration Procedures

The 418 candidates tested at five different sites in the state of California. At each site, experimental conditions were assigned to candidates in a spiral fashion. Each candidate took 150 items administered in a p&p mode and 150 in a Cadm mode. Every validation study item, like all bank items for the licensure program, could be presented on a single screen and did not contain graphics. Candidates had the capability to return and reconsider their answers to items. A total of approximately two and one-half hours was allotted for each of the two administration modes, allowing two hours of actual testing time for each mode after instructions and a Cadm practice session. In a traditional p&p administration, 325 items (including field test items) are given to candidates in a span of four hours.

MicroCAT software (1989) was used to present the Cadm items on IBM PS/2 Model 30 and Model 50 computers at the first four sites; Model 60 computers were used at the fifth site. All Cadm testing stations had IBM VGA color monitors.

Analyses

The configuration of the forms within experimental conditions permitted a step-wise series of independent tests of effects. Following a preliminary assessment of whether the spiralling of conditions produced equivalent samples, tests of the presence of an order of administration effect, an item allocation effect (fixed vs. random), and finally a mode of administration effect were performed on both scores and item parameters (one parameter difficulties or b-values). The presence of equivalent samples allowed the pooling of samples over conditions to attain more powerful tests of effects when previous tests in the series were insignificant.

Results

Equivalency of Samples

Table 1 contains summary statistics on raw scores obtained on the 90 item ABC form in each experimental condition. The range in mean raw scores of 57.03 to 58.77 correspond to a range of mean percent correct of 63.4% to 65.3%. These performance levels were very similar to the 64% to 66% average percent correct obtained on the last five p&p examinations. The two most extreme means mentioned above do not differ by more than twice the smaller of the two corresponding standard errors of the means (s.e.(mean)), which is .95 for Condition 2.

The raw score ranges as well as standard deviations (s.d.) shown in Table 1 also were very similar across conditions. The variances of the raw scores were equivalent across conditions when assessed by Cochran's Test ($c=.267$, ns at $p=.05$). There were no incidences of omitting throughout the ABC form for any condition.

There were, however, omissions in the DE or IJ forms positioned at the end of the p&p and Cadm administrations. The omit rate was higher for the Cadm items than the p&p, although for only one item was the omission rate over 1%. Two percent of the candidates omitted this Cadm item.

A strict test of sample equivalence would only compare scores from Condition 1 with scores from Condition 3 and similarly scores from Condition 2 with scores from Condition 4. Conditions within each of these two pairs had the same order of administration of the ABC form. The ABC items were the first 90 items administered in Conditions 1 and 3, the 151st through 240th items administered in Conditions 2 and 4.

However, the fact that all four samples did not have significantly different means or variances and that candidates were assigned to

conditions through spiralling suggests that the samples for Conditions 1 and 3 may be assumed to be equivalent to those for Conditions 2 and 4. The spiralling of candidates to conditions makes it unlikely that candidates in Conditions 2 and 4 happen to differ from those candidates in Conditions 1 and 3 by a constant amount that was exactly compensated for by an effect of late administration in the opposite direction.

Order of Administration Effect: b-values

Since there were no indications that the four samples were not equivalent, the samples for Conditions 1 and 3 were pooled (and called Conditions 1+3), as were the samples for Conditions 2 and 4 (Conditions 2+4). The ABC form was calibrated in each of the aggregated groups, the b-value for each item was paired up across orders of administration, and summary statistics were compiled for the 90 b-value differences.

Additionally the items in Form FGH were calibrated in Condition 1 and again in Condition 2, b-values were once again paired up, and summary statistics were compiled on these differences. Conditions 3 and 4 were not used in the calibration of Form FGH because of the possibility (at this point unassessed) of an effect due to the random administration of the G items. Both sets of summary statistics are presented in Table 2.

The two mean differences for the ABC pairs, with b-values calibrated on more than 200 candidates, and the FGH pairs, based on approximately 100 candidates, do not differ significantly from 0 when assessed by t tests on paired differences ($t = -.84$ and $.20$ with probabilities of .40 and .85, respectively). Both difference samples were unimodal and normally distributed by the Kolomogorov D statistic ($Pr > D = .07$ and $.10$, respectively). Figure 2 portrays the b-values for the ABC calibration in Conditions 1+3 plotted against the ABC b-values from Conditions 2+4. The relationship was decidedly linear ($r = .95$). Figure 3 illustrates the linear relationship between FGH values for Condition 1 versus Condition 2 ($r = .95$). Neither plot suggests that differences in b-values for the items in Form ABC and Form FGH may be attributed to broad positioning or context effects such as administration before rather than after a sequence of 150 items. The b-value differences may be more plausibly attributed to sampling or random estimation error.

Order of Administration Effect: Scores

IRT one-parameter score estimates (thetas) were obtained from the calibrations of Form ABC in Conditions 1+3 and Conditions 2+4 and Form FGH in Condition 1, then Condition 2. Because LOGIST standardizes the ability or theta distribution to have a mean of 0 and a standard deviation of 1.0 in each calibration run, it was

necessary to free each distribution of scores from this scaling constraint. In order to accomplish this, the set of b-values in each calibration run was transformed to have a mean of 0 and a standard deviation of 1.0. These slope and intercept transformation constants were then applied to the thetas estimated in that calibration run.

Summary statistics were obtained for the two sets of ABC scores and the two sets of FGH scores and are presented in Table 3. The difference of .012 between the two mean ABC thetas was insignificant ($t = .19$, the probability of attaining a t larger in absolute value by chance alone (significance probability) was .85). The two ABC standard deviations also were not significantly different (folded F or $F' = 1.06$, $p = .67$; Steel and Torrie (1980)). Similarly, the FGH means differed by an insignificant -.041 and the standard deviations also were not significantly different ($F' = 1.10$, $p = .63$). As with the b-values and raw scores, the thetas do not suggest the presence of an order of administration effect. Consequently they also serve to confirm the equivalence of samples across conditions.

Effect of Fixed vs. Random Item Allocation: b-values

The absence of an order of administration effect on b-values permitted a pooling of Conditions 1 and 2 and Conditions 3 and 4 for the purpose of obtaining larger calibration samples for a test of the equivalence of b-values across type of item allocation. After unscrambling to a common sequence the set of responses to the G set of items from the candidates in Conditions 3+4, the 30 items were calibrated twice, their b-values paired up, and summary statistics compiled on the differences. Table 4 contains these summary statistics.

The mean of the paired differences, .00, was of course non-significant ($t = 0.0$). The standard deviation of the differences, .24, was small relative to the average b-value standard error in each calibration group. The strong linear relationship between the two sets of b-values ($r = .97$) is evident in Figure 4.

The unimodal nature of the distribution of differences and their normality (Shapiro-Wilk Statistic (W) = .97, $p = .54$) strongly implies the absence of an effect of random item location. The b-values for the sample of 30 G items did not manifest signs of an immediate context effect, that is an effect on candidate performance due to the items immediately preceding and following an item, at the level of the population of short sequences of items.

Effect of Fixed vs. Random Item Allocation: Scores

Table 5 contains summary statistics for the thetas and raw scores generated using the G item set and the two calibration samples

mentioned above. Each theta distribution was freed from the LOGIST scaling constraints in the manner specified in the section called "Order of Administration Effect: Scores" above.

The theta means differed by a nonsignificant .030 and the raw score means by a nonsignificant .1. The pairs of standard deviations within each type of score were nearly identical. The absence of a significant difference between the two theta and the two raw score means suggests that theoretical predictions of test performance at the population level were not biased because of the random assignment of item location.

Mode of Administration Effect: b-values

Forms DE and IJ were used to examine differences between p&p and Cadm administration procedures. Form DE was calibrated first in Conditions 1+2 where it was administered in a p&p mode, then in Conditions 3+4 in a Cadm mode. The two b-values for each item were then paired and differences obtained. Similarly Form IJ was calibrated in Conditions 3+4 as p&p and Conditions 1+2 in a Cadm mode. b-value difference were obtained for both sets of pairs (Cadm - p&p). Summary statistics on DE and IJ b-value differences are presented in Table 6.

Both mean differences were not significantly different from 0. The DE and IJ difference means of -.05 and .02, respectively, do not even suggest a systematic, if insignificant effect, of one mode over the other as the two mean differences are in opposite directions. The standard deviations of differences were almost equivalent across difference samples (.22 vs. .24). The correlation between pairs of DE b-values and IJ b-values was very high; .97 for DE and .96 for IJ. Plots of both sets of b-value pairs are presented in Figures 5 and 6. Both distributions of differences were unimodal and normal by Kolomogorov D statistics (Pr. > D > .15 for both distributions).

The p = .10 significance level noted in Table 6 for the mean DE difference prompted a further examination of these differences to see if this marginally insignificant p value could be attributed to unique characteristics of at least some DE items. There was no sign that a few extreme differences were responsible. Under the assumption of normality, .6 DE differences would be expected to be more extreme than the mean, plus or minus three standard deviations. Three differences could be expected to be more extreme than two standard deviations. No difference and three differences actually were obtained, respectively.

There was also no indication that either the entire set of D item differences or the set of E item differences was extreme as a group. Both the D and E mean differences did not differ significantly (at p = .05 or .10) from 0, and each set of differences was approximately normally distributed. The D pairs were correlated .96 across modes; the E pairs were correlated .93.

A second check on the effect of computer administration on b-values consisted of a comparison of Cadm b-values with corresponding b-values obtained from the last p&p administration of the items. The ABCDE and FGHIJ forms each were calibrated in Conditions 1+2, and the resultant item b-values paired with the items' last p&p b-values. The FGHIJ form provided an examination of the Cadm mode while the ABCDE form (p&p) was used as a control. The mean difference for each form of 150 items and the correlation across modes are presented in Table 7. Figure 7 depicts the relationship between the Cadm FGHIJ form and the last p&p administration of these items.

Identical mean differences of .10 and nearly identical correlations across modes (.933 vs. .934 for FGHIJ and ABCDE, respectively) strongly suggest no effect of computer administration on b-values. It should be noted that the non-zero mean b-value difference of .10 represents an equating constant for the two modules since no attempt was made to equate these two forms or modules to the item bank from which they were selected. When contrasted to the average of the last five p&p equating constants, -.02, the difference of .10 implies the validation study candidates performed a little more poorly than did the candidates last taking the validation items in a p&p exam. Performance over the last five p&p exams tended, however, to be a little better on exam items compared to when these exam items were previously administered.

Finally Table 8 contains product moment correlation coefficients for various validation study Cadm and p&p b-value comparisons. Correlations of either Cadm vs. Cadm or Cadm vs. p&p validation b-values may be compared against correlations of validation Cadm and last p&p b-values. Where more than one correlation coefficient was available for one of the two types of comparisons at a particular combination of sample size and number of items, the smallest available correlation was recorded in the table.

It can be seen that the relationship between Cadm b-values and either other validation study Cadm or p&p b-values (in boldface) was very similar to the relationship of validation Cadm b-values to last p&p b-values. All correlations are high, even those based on small samples of approximately 100. Correlation coefficients for pure validation comparisons tended to be slightly higher than those involving last p&p b-values, which may perhaps be attributed to factors such as differential scale drift over time. Cadm correlations were also very similar to pure p&p correlations. For example, the correlation between Form ABC administered in Conditions 1+3 vs. Conditions 2+4 was .95. The correlation between Form ABCDE administered in Conditions 2+4 and the corresponding last p&p b-values was .93.

Mode of Administration Effect: Scores

Forms ABCDE and FGHIJ were each calibrated in Conditions 1+2 and thetas obtained for each form for each candidate. The theta distributions were once again freed from the scaling constraints

by using DE and IJ means and standard deviations to fix the item difficulty scale within each condition by mode of administration combination. Differences between Cadm (form FGHIJ) and p&p (Form ABCDE) scores were then obtained. A parallel procedure was performed on Forms ABCIJ and FGHDE in Conditions 3+4. Summary statistics are provided for both distributions of score differences in Table 9.

The two distributions of score differences were similar. Both had mean theta differences which were not significantly different from 0 and identical standard deviations of .35.

Raw scores on the DE form were compared across Conditions 2 (p&p) and 3 (Cadm) for signs of a mode effect in the raw score metric. The DE items are the last 60 items administered in these two conditions. Both mean raw scores and standard deviations were virtually identical across conditions. Condition 2 had a mean of 24.21 with a standard deviation of 5.14. Condition 3 had a mean of 24.23 and a standard deviation of 5.02.

Also included in Table 9 are summary statistics for complete validation study distributions of Cadm and p&p thetas (n = 418). A plot of these scores is shown in Figure 8. Once again, both distributions of scores (as opposed to the score differences above) were very similar. The difference of -.021 between the means was not significant when compared against either of the near equivalent mean standard errors. Distribution standard deviations were very similar (.645 vs. .681 for Cadm and p&p respectively), as were the interquartile ranges (Q3-Q1 was .968 for Cadm and .939 for p&p). The very similar skewness and kurtosis coefficients confirm the similar distribution shapes. The p&p distribution was normal by the Kolomogorov D statistic though the Cadm distribution was marginal (p = .05). Thetas based on these 150 item forms correlated .865. Given the absence of a mode of administration effect, this may be construed as a parallel forms reliability coefficient.

Table 10 contains the classifications when the 418 Cadm and 418 p&p thetas from Table 9 were scored as pass-fail after the cutscore was appropriately set.

The passing rates of 70.6% for the Cadm and 73.9% for the p&p administrations were very similar to the average 72.6% total operational p&p pass rate (computed over the last five administrations). The Cadm proportion passing did not differ significantly from the validation p&p proportion passing when assessed by McNemar's test for correlated proportions (X^2 = 3.13, ns at p = .05). Of the 418 validation study candidates, 87.1% were concordantly classified (pass or fail) by the two modes of administration.

The 87.1% concordance rate, based on the short Cadm and p&p forms each of which was 150 items long, must be considered as a lower bound to what would be obtained for the Cadm operational form which would be 230 items, or more than one and one-half times the length of the short forms. A rough estimate of what a concordance rate

would be for a 230 item Cadm operational exam may be obtained through the reliability coefficients and concordance proportions presented in Table 11.

Reliability coefficients for both the Cadm and p&p forms of varying lengths in Table 11 include coefficient alpha, an index of internal consistency, parallel forms reliability (i.e., correlation of scores across the pair of Cadm and p&p forms), and concordance rates based on the same cross-mode pair of forms. The increase in reliability that comes with increasing form length is plainly visible in alphas that increased from .81 (Cadm) for the 60 item Form IJ to .93 for the 240 item, predominantly p&p form, ABCDEFGH. The .93 alpha for the ABCDEFGH form, which was 10 items longer than an operational Cadm form, was very similar to the .92 to .95 alphas typically obtained for p&p exams. This similarity of Cadm and p&p reliabilities is also evident in the near equivalence of alphas for the smaller form lengths listed (e.g., .84 for Cadm FGH, .83 for p&p ABC).

The change in reliability with increasing test length is also evident in parallel forms and concordance indices. Though slightly below the alphas, both indices increase when test length increased from a 90 item to a 150-item form 1.7 times as long. A concordance proportion of .84 for the 90-item forms, which was attended by a nonsignificant (at $p = .05$ or .10) McNemar's test of differences in proportion passing across modes, increased to .87 for 150-item forms. A conservative, extrapolated concordance rate for a 230 item Cadm form might be presumed to be at least .89.

SUMMARY and CONCLUSIONS

The following conclusions may be drawn from the statistical analyses described above.

(1) An assessment of the efficacy of spiralling to produce equivalent samples across experimental conditions produced no indication that candidates were differentially motivated across administration modes or in performance on the validation items relative to previous p&p examination performance.

(2) Forms that were constructed to be a priori parallel across conditions were verified to be parallel across the two modes of administration, suggesting no difference in reliability due to administration mode.

(3) There was no sign of an effect of random assignment of item location on thetas, raw scores, or item b-values. These results suggest that there is no evidence for the presence, at the population level, of immediate context effects on item performance (i.e., an unpredicted effect on item performance due to the items that immediately surround a given item).

(4) The absence of order of administration effects on several different forms of items does not support the presence of

positioning effects over longer sequences of items. These findings are not unexpected due to the nature of the exam items: single screen, no graphics, and not heavily dependent upon reading comprehension.

(5) Mode of administration effects were not found for item b-values or scores when assessed by powerful tests of paired differences. Cadm b-values were highly correlated with past p&p b-values. Score distributions based on forms shorter than the Cadm operational forms were very similar with corresponding p&p forms. There were no signs of significant differences in overall performance levels or individual pass-fail decisions across modes.

(6) The nature of the design permitted assessment on multiple item forms and candidate samples. This test redundancy reinforced the basic conclusion that there is no significant effect of computer administration on candidate performance.

REFERENCES

American Psychological Association (1986). Guidelines for computer-based tests and interpretations. Washington, D.C.: American Psychological Association.

Assessment Systems Corporation (1989). User's manual for the MicroCAT testing system. St. Paul, MN: C. David Vale.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.), Educational Measurement (3rd ed., pp. 367-407). New York: National Council on Measurement in Education.

Mazzeo, J. & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests. College Board Report No. 88-8. New York: The College Board.

Steel, R.G.D., & Torrie, J.H. (1980). Principles and procedures of statistics. New York: McGraw-Hill.

Wainer. H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 3, 185-201.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 2, 245-262.

Table 1

Mean Raw Scores on Form ABC for the Four Experimental Conditions

	Condition			
	1	2	3	4
Mean Raw Score	58.77	57.03	57.66	58.42
Mean % Correct	65.3	63.4	64.1	64.9
n	108	105	103	102
Min./Max.	28/79	34/74	38/82	33/77
s.d.	10.17	9.69	9.79	9.66
s.e.(mean)	.98	.95	.96	.96

Table 2

b-value Differences Obtained for Early and Late Administrations of Forms ABC and FGH

	# Items	Mean	Standard Deviation	Minimum Value	Maximum Value	t	Pr>|t|	Normal Pr > D
ABC (Conditions 1+3 vs Conditions 2+4)	90	-0.03	0.30	-0.99	0.71	-0.84	0.40	.07
FGH (Condition 1 vs Condition 2)	90	0.01	0.35	-0.81	0.96	0.20	0.85	.10

Table 3

Theta Summary Statistics for Orders of Administration

Form	Conditions	n	Mean	s.d.	Min.	Max.	t	Prob > |t|
ABC	1+3	211	.795	.626	-0.97	2.75	0.19	0.85
	2+4	207	.783	.608	-0.68	2.19		
FGH	1	108	.663	.605	-1.40	1.88	-0.50	0.62
	2	105	.704	.577	-0.45	2.02		

Table 4

b-value Differences Across Item Allocation

Index	n	Mean	s.d.	Min.	Max.	t	Pr > |t|
b-value difference	30	-.00	.24	-.50	.65	-0.00	1.00

Table 5

Difference in Scores on Item Set G
Across Type of Item Allocation

Conditions	n	Mean	s.d.	Min.	Max.	t	Pr >|t|
			Thetas				
1+2	213	.741	.858	-1.84	3.40	.365	.715
3+4	205	.711	.822	-1.27	3.23		
			Raw Scores				
1+2	213	18.7	4.17	6	28	.158	.874
3+4	205	18.6	4.20	8	28		

Table 6

Differences in b-values for the DE and IJ Items Administered
Across Mode of Administration (Cadm - p&p)

Form	n	Mean	s.d.	Min.	Max.	t	Pr > |t|
DE	60	-.05	.22	-.46	.50	1.69	.10
IJ	60	.02	.24	-.54	.58	.68	.50

BEST COPY AVAILABLE

Table 7

Comparison of b-values from the Validation Study with Last Previous Paper-and-Pencil Standard Administration

Form (mode)	Conditions	Items	Mean dif. from p&p	Corr. with last p&p
FGHIJ (Cadm)	1+2	150	.10	.933
ABCDE (p&p)	1+2	150	.10	.934
Average of last 5 p&p equating constants:			-.02	

Table 8

Correlation of b-values for Various-Sized Forms (P&P Forms, Both Samples)

	# Items			
Sample Size (app.)	30	60	90	150
100	.97	.92	.91, .95	.91
200		.94, .96	.93	.93
400			.94	

Bold face correlations have the same sample size for each group used to compute the correlation. Plain face correlations are against the last paper-and-pencil b-value based on approximately 200 candidates.

BEST COPY AVAILABLE

Table 9

Differences in Thetas for Paper-and-Pencil and Computer Administrations

Forms (mode)	Conditions	n	Mean	s.d.	Min.	Max.	t	Pr > |t|
		Differences						
ABCDE (p&p), FGHIJ (Cadm)	1+2	213	.04	.35	-1.03	1.44	1.61	.11
ABCIJ (p&p), FGHDE (Cadm)	3+4	205	.00	.35	-1.67	0.98	0.09	.93

Distributions (n = 418)

Mode	Mean	s.d.	Q3-Q1	s.e (mean)	Skew	Kurtosis	D (Normality)	Pr > D
Cadm	.598	.645	.968	.032	.028	-.478	.044	.05
p&p	.619	.681	.939	.033	.059	-.324	.033	>.15

Correlation: Cadm vs p&p: .865

Table 10

Computer Administered and Paper-and-Pencil Validation Study Classification Decisions

		P & P Pass	P & P Fail	
Cadm	Pass	275	20	295 (70.6%)
	Fail	34	89	123 (29.4%)
		309	109	418
		(73.9%)	(26.1%)	

Proportion Concordant Classifications: 87.1%

McNemar's Test of Difference in Proportion Passing:

$\chi^2 = 3.13$

ns at p = .05

BEST COPY AVAILABLE

Table 11

Reliability and Concordance Statistics for Cadm and P&P Forms

	Mode / Form				
	Cadm	IJ	FGH	FGHIJ	
	p&p	DE	ABC	ABCDE	ABCDEFGH*
Within Mode					
Coefficient Alpha		.81	.84	.90	
		.80	.83	.90	.93
Across Mode					
Parallel Forms			.81	.86	
Concordance			.84	.87	

* This form contains 150 p&p and 90 Cadm items.

Figure 1

Validation Study Design

Condition	1		2	
Order of Administration	1	2	1	2
Mode of Administration	P&P	Cadm	Cadm	P&P
Item Sets :	A B C D E	F G H I J	F G H I J	A B C D E

Condition	3		4	
Order of Administration	1	2	1	2
Mode of Administration	P&P	Cadm	Cadm	P&P
Item Sets :	A B C I J	F G* H D E	F G* H D E	A B C I J

Each letter denotes a set of 30 items

Boldface denotes items that are administered in both modes

* Denotes items whose locations within the G item set were assigned randomly

Figure 2

Plot of ABC b-values: Conditions 1+3 vs Conditions 2+4

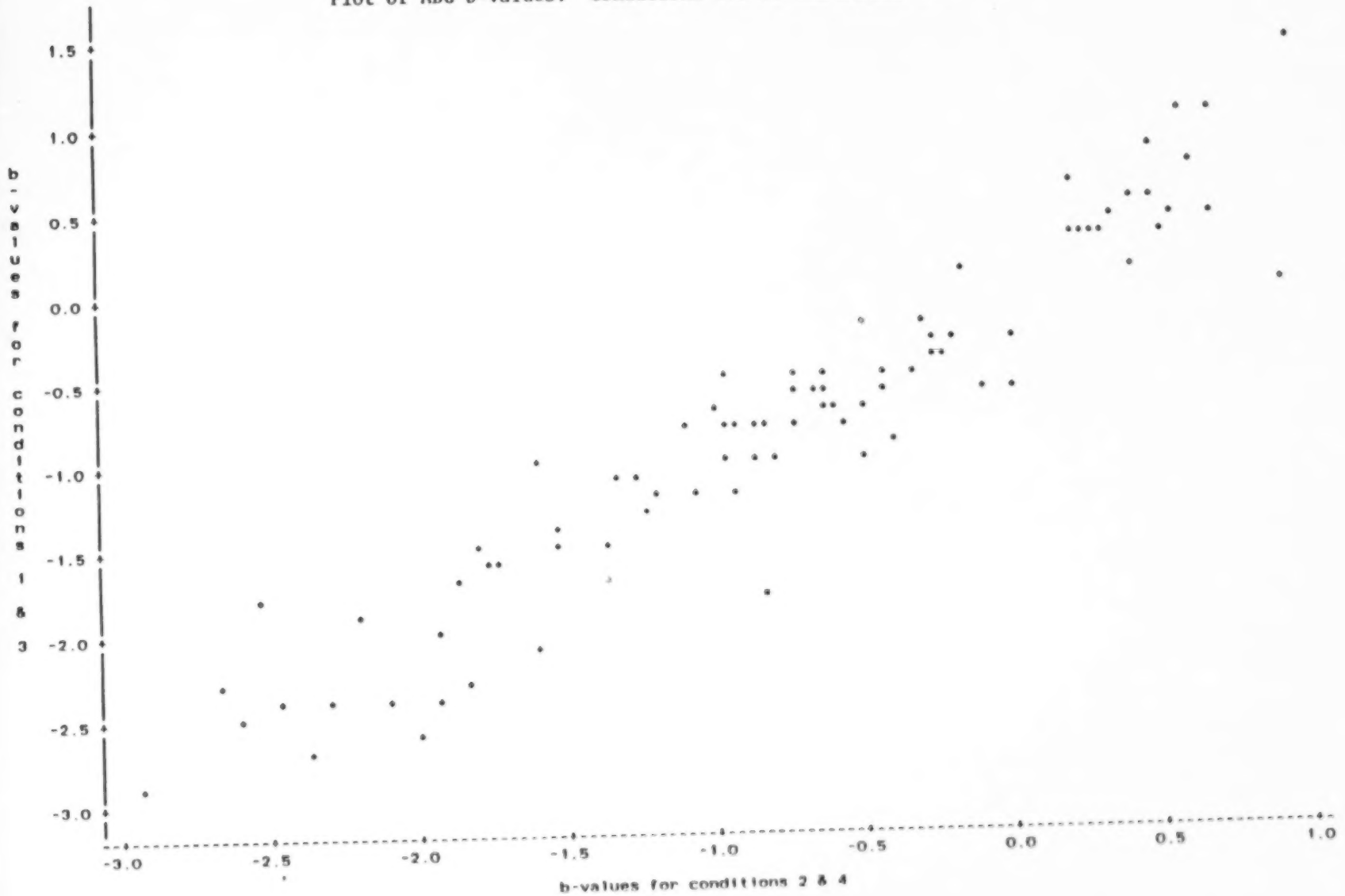

Figure 3

Plot of FGH b-values: Condition 1 vs Condition 2

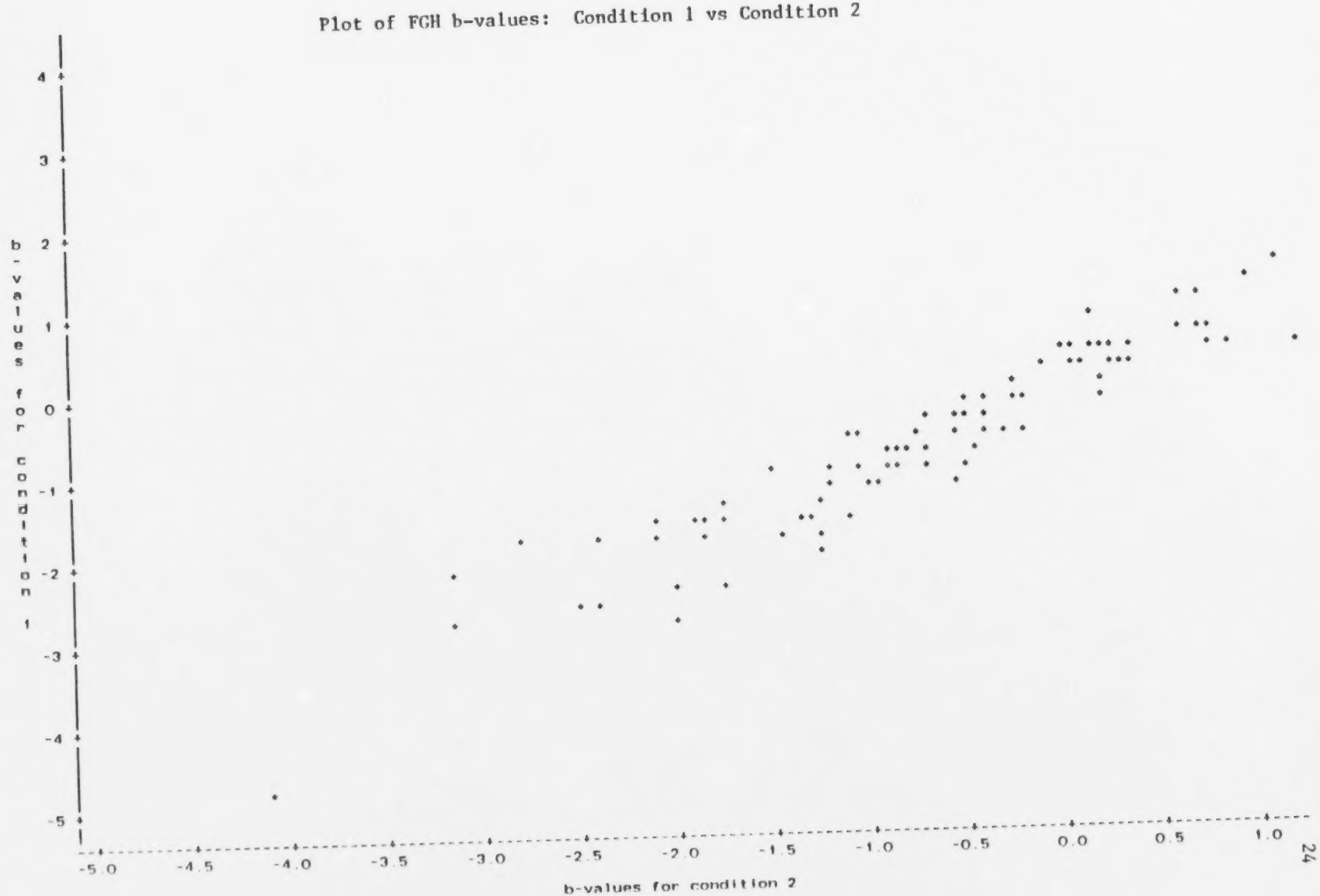

Figure 4

b-values for the Thirty G Items: Fixed vs Random Location

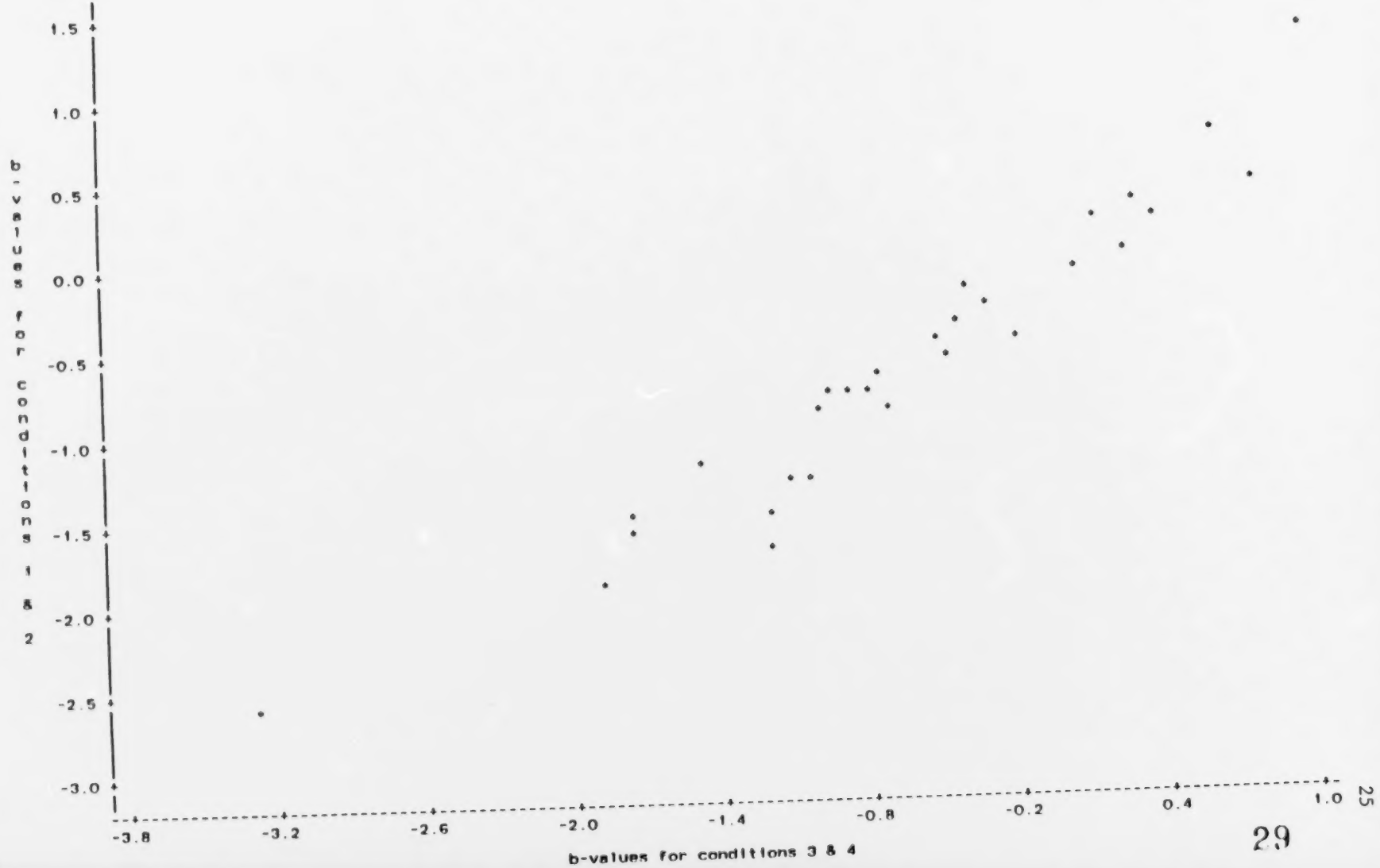

Figure 5

Plot of DE P&P b-values vs DE Cadm b-values

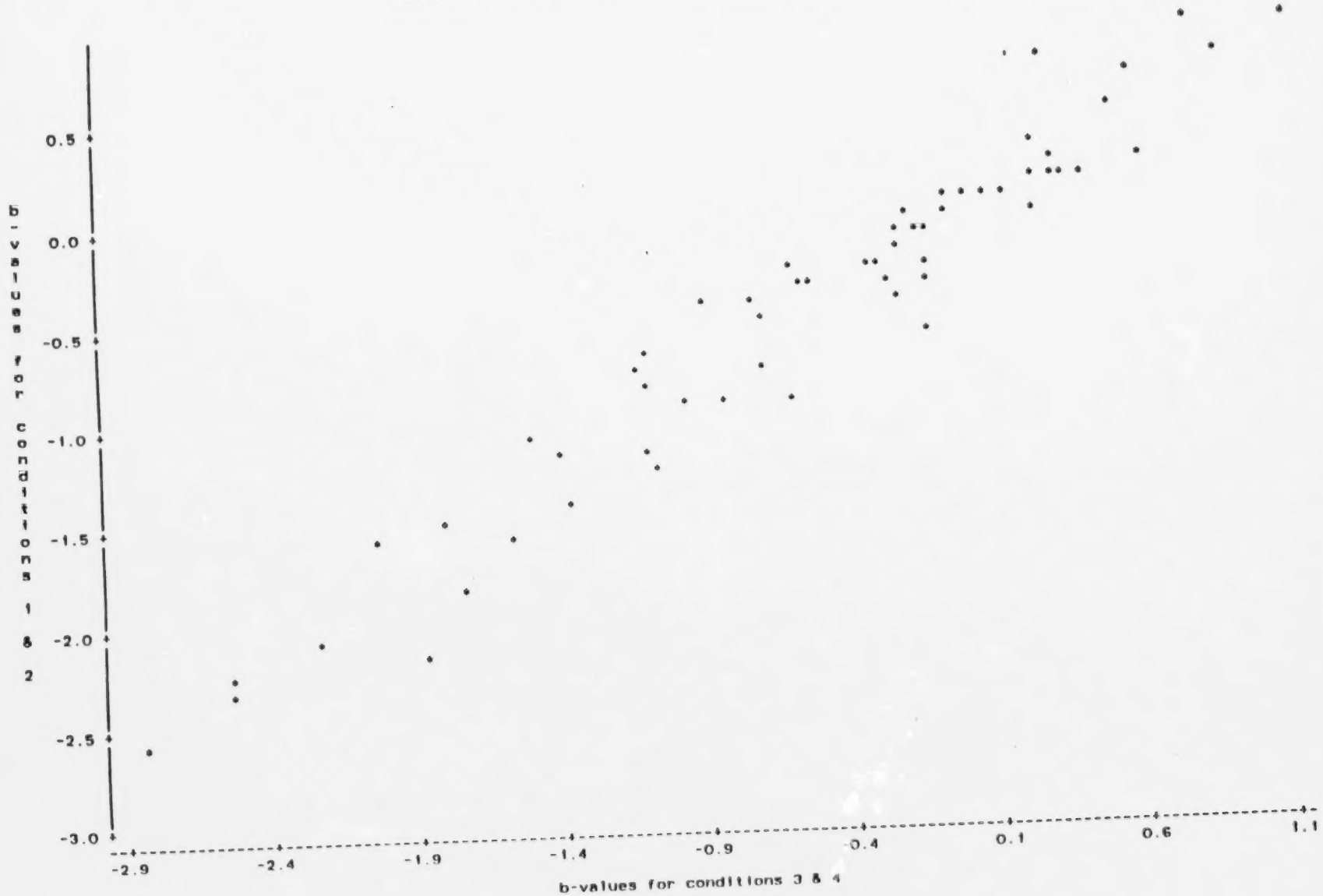

Figure 6

Plot of IJ P&P b-values vs IJ Cadm b-values

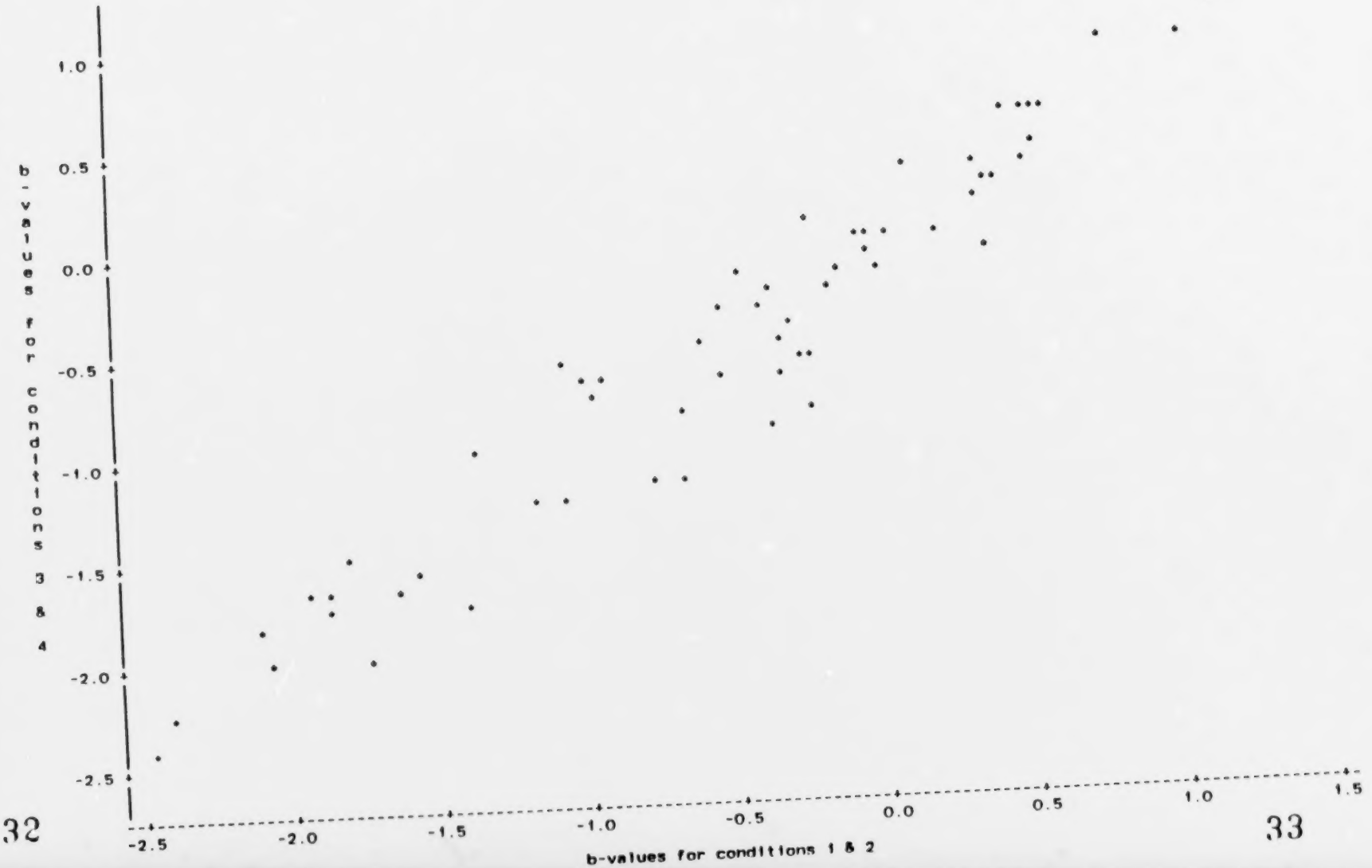

Figure 7

Plot of Last P&P b-values vs. Cadm Form FGHIJ b-values (Conditions 1+2)

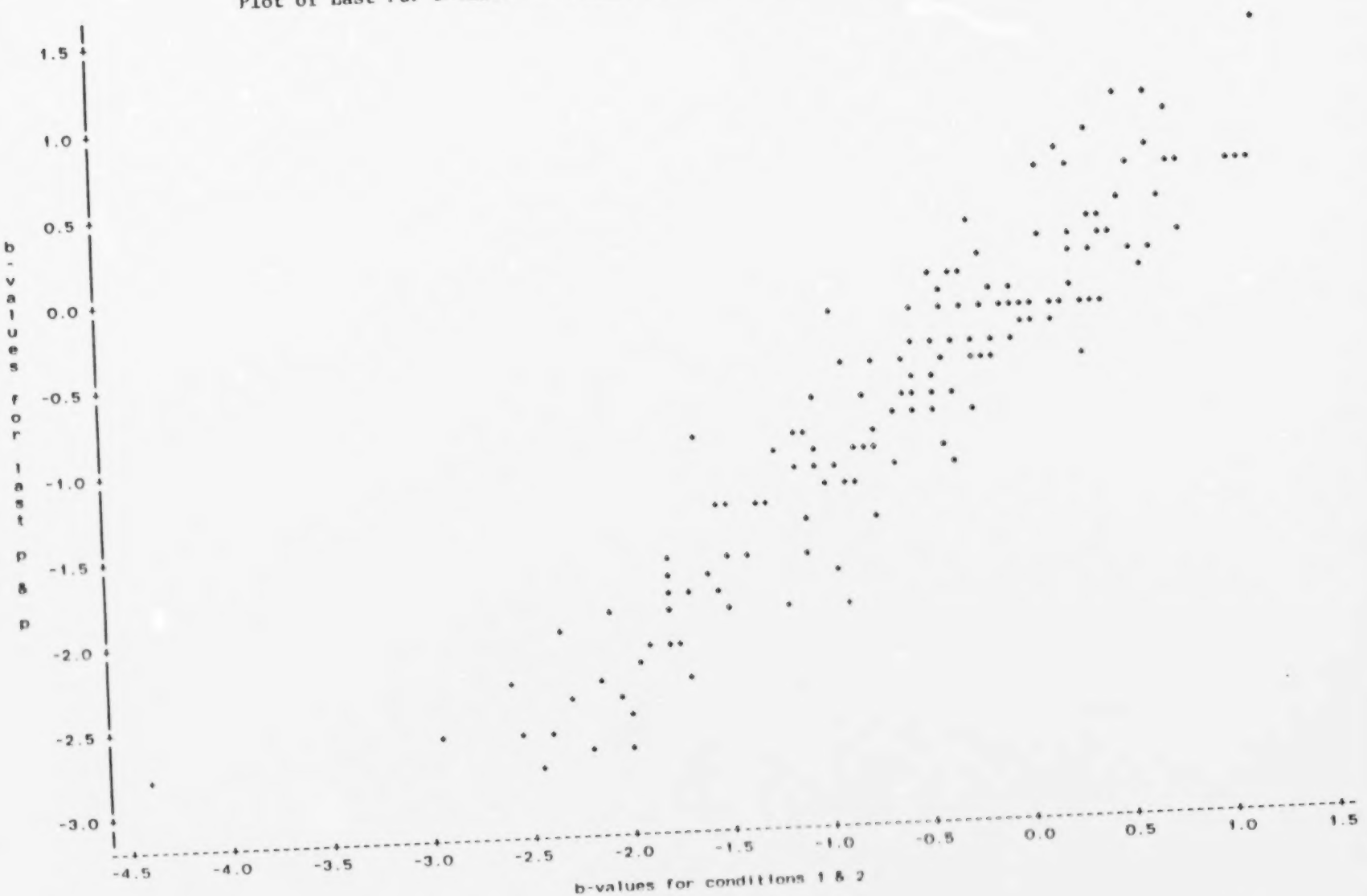

Figure 8

Plot of Thetas: P&P vs Cadm